

Ordered Decompositional DAG Kernels Enhancements

Giovanni Da San Martino^a, Nicolò Navarin^{b,*}, Alessandro Sperduti^b

^a*Qatar Computing Research Institute, HBKU, P.O. Box 5825 Doha, Qatar*

^b*Department of Mathematics, University of Padova, via Trieste 63, Padova, Italy*

Abstract

In this paper, we show how the Ordered Decomposition DAGs (ODD) kernel framework, a framework that allows the definition of graph kernels from tree kernels, allows to easily define new state-of-the-art graph kernels. Here we consider a fast graph kernel based on the Subtree kernel (ST), and we propose various enhancements to increase its expressiveness. The proposed DAG kernel has the same worst-case complexity as the one based on ST, but an improved expressivity due to an augmented set of features. Moreover, we propose a novel weighting scheme for the features, which can be applied to other kernels of the ODD framework. These improvements allow the proposed kernels to improve on the classification performances of the ST-based kernel for several real-world datasets, reaching state-of-the-art performances.

Keywords: Kernel Methods, kernel functions, Graph Kernels, Classification

1. Introduction

The increasing availability of data in structured form, such as trees [1] or graphs [2, 3, 4], has led to the development of machine learning techniques able to deal directly with such types of data. Among these, kernel methods, such as Support Vector Machines (SVM) [5], have become very popular due to their generalization ability and state of the art performances in many tasks, such as relationship extraction [6], analysis of RDF data [7], action

*Corresponding author

Email addresses: gmartino@qf.org.qa (Giovanni Da San Martino),
nnavarin@math.unipd.it (Nicolò Navarin), sperduti@math.unipd.it (Alessandro Sperduti)

recognition [8], text categorization of biomedical data [9] and bioinformatics [10].

The class of kernel methods comprises all those learning algorithms which do not require an explicit representation of the input, but only information about the similarities among them. A simple way of assessing the similarity between two objects described by a set of features is to compute the dot product of their representation in feature space. If a “similarity” function $k(\cdot, \cdot)$, corresponding to a dot product $\langle \cdot, \cdot \rangle$ in feature space, is available, the intermediate step of explicitly representing the data can be avoided. In fact, computing $k(x_1, x_2)$ implicitly corresponds to perform a nonlinear transformation of the input vectors x_1 and x_2 via a function $\phi(\cdot)$ and then to compute the dot product of the resulting vectors $\phi(x_1)$ and $\phi(x_2)$. The function $\phi(\cdot)$ projects the input vectors into a feature space of much higher (possibly infinite) dimension where it is more likely to accomplish the learning task. Kernel methods generally formulate a learning problem as a constrained optimization one, where an objective function combining an empirical risk term with a (quadratic) regularizer must be minimized. If the employed kernel function is symmetric positive semidefinite, the problem is convex and thus has a global minimum [5].

Any kernel method can be decomposed into two modules: *i*) a problem specific kernel function; *ii*) a general purpose learning algorithm (the solver). Since the solver interfaces with the problem only by means of the kernel function, it can be used with any kernel function, and vice-versa. Examples of popular kernel methods are the perceptron [11] for the on-line setting, and the Support Vector Machines [5] for the batch setting. Note that, provided an appropriate kernel function is given, any kernel method can be applied to any type of data. More importantly, the kernel function encodes all the information about the input data, thus the definition of appropriate kernel functions is crucial for the outcome of the learning algorithm.

A popular strategy for defining kernel functions for structured data is to decompose the structures into their constituent parts, and then, for each pair of parts, apply a local kernel [12]. While this strategy has been proved successful for strings and trees [13, 14, 15, 16, 17, 18], it is not directly applicable to graphs because of the computational complexity issues which arise: representing a graph in terms of its subgraphs is not feasible since subgraph isomorphism, an *NP-complete* problem, should be solved for each pair of subgraphs. In [19] it has been demonstrated that, any kernel whose feature space mapping is injective, is as hard to compute as graph isomorphism, an

NP problem that still is not known whether it is in P or if it is NP -complete. Due to this limitation, the available strategies for building kernels are: *i*) restricting the input domain to a class of graphs for which isomorphism can be checked quickly [20]; *ii*) select a priori a set of features, usually corresponding to a specific type of substructure, such as walks [19], paths [21, 22], subtree patterns [23, 24]. The former approach can be applied to a limited type of graphs, the latter tends to have a limited flexibility: when the available kernels are not relevant to the task, a new one has to be designed. However, defining an efficient symmetric positive semidefinite kernel, corresponding to the desired feature space, can be an extremely difficult task. All the above approaches discard information about the original graph and are effective only when the selected features are relevant for the current problem. We propose to design graph kernels as follows: first transform the graphs into simpler structures, i.e. multisets of directed acyclic graphs (DAGs), and then extend the definition of a large class of already available kernels for trees to DAGs. Our approach allows the application of the vast literature on kernels for trees, which consists of fast and/or very expressive kernels, to the graph domain.

Generally speaking, a serious drawback which prevents many of the kernels listed above to be applied to large datasets is their computational time complexity. Those kernels which can be applied to large datasets exploit a “limited” number of features to represent a graph. For example, the kernel proposed in [24] has a linear complexity in the number of edges of the graphs because any graph is represented in the feature space by a number of non-zero features which is proportional to the number of nodes of the graph. On the other hand, a too compact representation of a graph in feature space may have a negative impact on the effectiveness of the kernel because of a reduced discrimination ability.

In this paper, we tackle this problem by proposing various enhancements to a fast graph kernel based on the Subtree kernel for trees (ST) [25]. Among these, the main contribution is a novel tree kernel, which has the same worst-case complexity of the ST kernel, while the size of its feature space is much larger.

The paper is structured as follows. Section 2 introduces some basic notation and definitions. Section 3 recalls the ODD framework, of which the proposed kernels are instances. Section 4 describes the main contributions of the paper: the $ST+$ kernel for DAGs and a novel weighting scheme for the features, which can be applied to other kernels of the ODD framework.

Section 5 discusses some related kernels for graphs, and Section 6 provides experimental evidence of the effectiveness of the proposed approaches. Finally, Section 7 draws conclusions.

The paper extends the work in [26] by adding: *i*) a self-contained and simplified description of the $ST+$ kernel; *ii*) a novel, more effective, feature weighting scheme; *iii*) an extended and revised “Related Work” section; *iv*) a novel set of experiments which are now performed on much larger benchmark datasets and for a larger number of competing graph kernels; *v*) a comparison among empirical execution times of the various experimented kernels.

2. Notation

A graph is a triplet $G = (V, E, L)$, where V (alternatively V_G) is the set of nodes ($|V|$ is the number of nodes), E the set of edges and $L()$ a function returning the label of a node. All labels are obtained from a fixed alphabet \mathcal{A} . A graph is undirected if $(v_i, v_j) \in E \Leftrightarrow (v_j, v_i) \in E$, otherwise it is directed. A path in a graph is a sequence of nodes v_1, \dots, v_n such that $v_i \in V, 1 \leq i \leq n$, $(v_i, v_{i+1}) \in E$ and $\forall 1 \leq i \leq n, 1 \leq j < n, j \neq i. v_i \neq v_j$ (no node, except the first one, can appear twice in the same path). A cycle is a path for which $v_1 = v_n$; a cycle is even/odd if its number of nodes is even/odd, respectively. A graph is connected if there exists a path connecting each pair of nodes. A connected graph is rooted if exactly one node has no incoming edges. A graph is ordered if the set of neighbours of each node is ordered. A tree is a rooted connected directed acyclic graph where each node has at most one incoming edge. A subtree of a tree T is a connected subset of nodes of T . A proper subtree is a subtree composed by a node and all of its descendants. Given a node v of a tree, $\rho(v)$ represents the outdegree of v , i.e. the number of nodes connected to v . We will use ρ as the maximum outdegree of a node in either a tree or a graph. The depth $depth(v)$ of a node v is the number of edges in the shortest path between the root of the tree and v . If the tree is ordered, $ch_v[j]$ represents the j -th child of v and $chs_v[j_1, j_2, \dots, j_n]$ indicates the set of children of v with indices j_1, j_2, \dots, j_n . Given a graph G and a node $v \in V(G)$, we define a subtree-walk of size h as the tree obtained by the following procedure: the root of the tree is v ; at each step i , with $1 \leq i \leq h$, and for each current leaf node v_j of the tree, any neighbouring node of v_j in G is added to the tree as a child of v_j . Note that, when $h > 1$, typically a node of G can appear multiple times in the same subtree-walk. Given a DAG D and a node $v_i \in V(D)$, we define a tree-visit,

denoted by $\overset{v_i}{\Delta}$, as the tree resulting from the visit of D starting from the node v_i . Such visit returns all the nodes of D reachable from v_i . If a node v_j can be reached more than once, more occurrences of v_j will appear in $\overset{v_i}{\Delta}$ (see Figure 2-b for an example).

3. Preprocessing: from Graphs to Multisets of DAGs

This section recalls the ODD-Kernels framework for graphs [27]. The idea of our approach is to transform the graphs into simpler structures, i.e. DAGs, and then apply a kernel for such structures. The following subsections explain each step of the transformation.

3.1. From Graph to DAGs

The graph G is mapped into a multiset of DAGs $DD_G = \{DD_G^{v_i} | v_i \in V_G\}$, where $DD_G^{v_i} = (V_G^{v_i}, E_G^{v_i}, L)$ is obtained by keeping each edge in the shortest path(s) connecting v_i with any $v_j \in V_G$. From a practical point of view, $DD_G^{v_i}$ can be built by performing a breadth-first visit on the graph G starting from node v_i and applying the following rules:

1. during the visit a direction is given to each edge; if v_j is reached from v_i in one step, then $(v_i, v_j) \in E_G^{v_i}$ (note that edge (v_j, v_i) is not added to $E_G^{v_i}$);
2. edges connecting nodes reached at level l of the visit to nodes reached at level $g < l$ are not added to $E_G^{v_i}$ (such edges would induce a cycle in $DD_G^{v_i}$.)

For every choice of G and v_i , a single *Decompositional Dag* $DD_G^{v_i}$ is generated. By repeating the procedure for each node of the graph, $|V|$ DAGs are obtained. Figure 1 shows the four DD s obtained from the undirected graph in Figure 1-a. Note that when the same node is reached simultaneously (at the same level of the visit) from different nodes, then all involved edges are preserved. For example, when considering the visit at level 2 starting from node **s**, the node **d** is reached simultaneously by edges **(b, d)** and **(e, d)**, and both of them are preserved in the corresponding Decompositional DAG (see Figure 1-b). In order to reduce the total number of nodes of DD_G^v , we propose to limit the depth of the visits during the generation of the multiset of DAGs [27] to a constant value h . The resulting DAG will be referred to as $DD_G^{v,h}$. Given $v \in V_G$, let H be the number of nodes generated by the visits up to depth h . An upper bound for H is ρ^h . Notice, this is a loose bound,

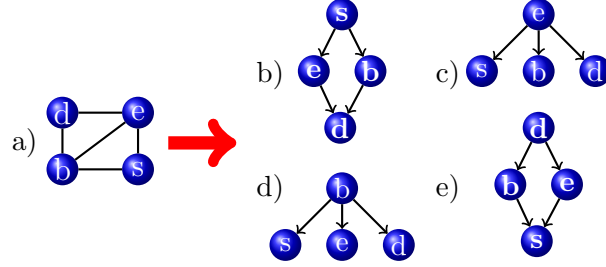


Figure 1: Example of decomposition of a graph a) into its 4 DDs b-e).

in many practical cases. The total number of nodes of DD_G is $|V_G|H$. Note that, if ρ is constant, then also H is constant.

3.2. Ordering DAG nodes

The kernels we define in the following, which are all straightforwardly derived from tree kernels, require DAG nodes to be ordered. Therefore, we define a strict partial order \prec between DAG nodes in $DD_G^{v_i}$ obtaining Ordered DAGs $ODD_G^{v_i}$. The ordering makes use of a unique representation of subtrees as strings inspired by [14]. Here we modify such mapping by employing perfect hash functions, i.e. hash functions which guarantee to have no collisions, to encode subtrees [24, 28]. Let $\kappa()$ be a perfect hash function, $\#, \lceil, \rfloor$ be symbols never appearing in any node label and $ch_v[j]$ the j -th node in the ordered sequence of outgoing edges of v , then

$$\pi(\overset{v}{\Delta}) = \begin{cases} \kappa(L(v)) & \text{if } v \text{ is a leaf node} \\ \kappa(L(v) \lceil \pi(\overset{ch_v[1]}{\Delta}) \# \dots \# \pi(\overset{ch_v[\rho(v)]}{\Delta}) \rfloor) & \text{otherwise} \end{cases} \quad (1)$$

where the children of v are recursively ordered according to their $\pi()$ values. To simplify notation, in the following, when it is clear from the context, we will use the notation $\pi(v)$ instead of $\pi(\overset{v}{\Delta})$. Then $v_i \prec v_j$ if $\pi(v_i) < \pi(v_j)$, where $<$ is the relation of order between alphanumeric strings. Notice that $\pi(v_i) = \pi(v_j) \Leftrightarrow \neg(v_i \prec v_j) \wedge \neg(v_j \prec v_i)$, i.e. $\pi(v_i) = \pi(v_j)$ if and only if the nodes v_i and v_j are not comparable. In such case, many orderings for non comparable children nodes according to \prec are possible. We are now going to prove some results that will make it easier to show, in Section 4, that each kernel described in this paper (as well as for a large class of kernels for trees) yield the same features, independently of the ordering of non comparable nodes. Since all the features of the kernels in Section 4 are extracted from

tree visits of DAG nodes, our goal here is to show that isomorphic DAGs yield the same tree visits. We first show that if two DAGs $DD_{G_1}^{v_i}$ and $DD_{G_2}^{v_j}$ are isomorphic, then the root nodes of the DAGs are not comparable with respect to the ordering $\dot{<}$, in fact:

Theorem 3.1. *if two DAGs $DD_{G_1}^{v_i}$ and $DD_{G_2}^{v_j}$ are isomorphic, then $\neg(v_i \dot{<} v_j) \wedge \neg(v_j \dot{<} v_i)$.*

Proof Let $f : V_{G_1} \rightarrow V_{G_2}$ be an isomorphism between $DD_1^{v_i}$ and $DD_2^{v_j}$. We prove the thesis by induction. Let $f(v_i) = v_j$, since the nodes are isomorphic $L(v_i) = L(v_j)$. If v_i and v_j are leaf nodes, then $\pi(v_i) = \pi(v_j)$ and consequently $\neg(v_i \dot{<} v_j) \wedge \neg(v_j \dot{<} v_i)$. Otherwise, by inductive hypothesis $\forall l. 1 \leq l \leq \rho(v_i)$. $\pi(ch_{v_i}[l]) = \pi(ch_{f(v_i)}[l])$ and $L(v_i) = L(f(v_i))$, thus $\pi(v_i) = \pi(f(v_i)) = \pi(v_j)$.

The following theorem shows that two non comparable nodes v_i, v_j , yield identical tree visits $\overset{v_i}{\Delta}, \overset{v_j}{\Delta}$:

Theorem 3.2. *Given the ordering $\dot{<}$, $\neg(v_i \dot{<} v_j) \wedge \neg(v_j \dot{<} v_i)$ if and only if $\overset{v_i}{\Delta}$ and $\overset{v_j}{\Delta}$ are identical.*

Proof If $\neg(v_i \dot{<} v_j) \wedge \neg(v_j \dot{<} v_i)$ then $\pi(v_i) = \pi(v_j)$. Recalling that $\kappa()$, the function on which $\pi()$ is based on, is a perfect hash function, we prove the thesis by induction. If v_i, v_j are leaf nodes, then $\pi(v_i) = \pi(v_j) \Leftrightarrow L(v_i) = L(v_j)$. If v_i, v_j are not leaf nodes, then $\forall l. 1 \leq l \leq \rho(v_i)$ $\overset{ch_{v_i}[l]}{\Delta}$ is identical to $\overset{ch_{v_j}[l]}{\Delta}$ for inductive hypothesis, and then it must be $L(v_i) = L(v_j)$ since $\pi(v_i) = \pi(v_j)$; therefore $\overset{v_i}{\Delta}$ is identical to $\overset{v_j}{\Delta}$. Now we show that if $\overset{v_i}{\Delta}$ is identical to $\overset{v_j}{\Delta}$, then $\pi(v_i) = \pi(v_j)$ by induction. The base case has already been proved by the equality $\pi(v_i) = \pi(v_j) \Leftrightarrow L(v_i) = L(v_j)$. By inductive hypothesis $\pi(ch_{v_i}[m]) = \pi(ch_{v_j}[m])$ for each child m of v_i and v_j . Then $\pi(v_i) = \pi(v_j)$ and $\neg(v_i \dot{<} v_j) \wedge \neg(v_j \dot{<} v_i)$.

Note that, since any ordering between non comparable vertices is equivalent for our goals, we avoid to give a specific ordering. If the $\pi()$ values are computed according to a post order visit of the DAG, then the values $\pi(ch_v[l])$ for $1 \leq l \leq \rho(v)$ are already available when computing $\pi(v)$. Thus the time complexity of the ordering phase of the DAG is $O(|V_G| \rho \log \rho)$ where the term $\rho \log \rho$ accounts for the ordering of the children of each node.

3.3. Applying Tree Kernels to DAGs

If we restrict to the kernels which are going to be presented in this paper, the general formula for graph kernels derived from the ODD framework [27] can be simplified as follows

$$ODD_K(G_1, G_2) = \sum_{\substack{D_1 \in ODD_{G_1} \\ D_2 \in ODD_{G_2}}} \langle \phi^K(D_1), \phi^K(D_2) \rangle, \quad (2)$$

where $\langle \cdot, \cdot \rangle$ is the dot product operator, and $\phi^K(D)$ is the explicit feature space projection of the DAG D with respect to the kernel K and $ODD_G = \{ODD_G^{v,h} | v \in V_G\}$. Section 4.1 gives an example of an instance of the kernel defined in (2).

4. Kernels for DAGs

In Section 3, we showed a preprocessing procedure for transforming a graph into a multiset of ordered DAGs. In this section, we first recall the ODD_{ST_h} kernel, presenting it in a slightly different way than as it was originally introduced in the paper [27]. Then, we describe the original contributions of the paper, i.e. a novel kernel for DAGs, named $ST+$, and a novel weighting scheme for the features which is specifically designed for our setting.

4.1. ST kernel for DAGs

Let us consider $\overset{v}{\Delta}$, the tree resulting from the visit of ODD_G^v starting from the root node v . The visit can be stopped when the tree $\overset{v}{\Delta}$ reaches a maximum depth h . Such tree is referred to as $\overset{v}{\Delta}|_h$.

As an example of kernel in (2), we recall the ODD_{ST_h} kernel [27]. The features of the kernel are $\overset{v}{\Delta}|_l$, for each $v \in V_D$, where $D \in ODD_G$ as defined in the previous section and for each $0 \leq l \leq h$. Specifically, any node v of the DAG contributes to the feature vector $\phi(\cdot)$ as $\phi_{\pi(v)} = \lambda^{\frac{size}{2}}$, where $size = |\overset{v}{\Delta}|_l$ for some l , and $\pi(v)$ (we recall that this notation stays for $\pi(\overset{v}{\Delta}|_l)$) is the function defined by (1). This weighting scheme for the features is inherited by the ST [25] kernel and it is motivated by the fact that when computing a kernel involving two matching large trees, the value returned by the kernel is very large because not only the whole trees will match, but

all their subtrees will match as well. To correct that, the contribution to the kernel of a matching tree is down-weighted by $\lambda^{\frac{size}{2}}$, where $0 < \lambda \leq 1$.

In order to demonstrate that the resulting graph kernel is positive semidefinite, we need to prove that our $\phi(\cdot)$ function is well-defined, i.e. it gives the same result when the representation of the input is changed without changing the value of the input. If two graphs are isomorphic, they generate the same multiset of DAGs (since they are defined over shortest paths). We know from Theorem 3.1 that isomorphic DAGs generate the same visits. Since the features considered by the ST kernel are subtrees, it directly follows from Theorem 3.2 that the swapping of non comparable vertices in the ordering do not affect the feature space representation of a graph. Thus, we provided a well-defined feature space representation for ODD_{ST_h} , from which it follows that the kernel is positive semidefinite.

4.2. The $ST+$ Kernel for DAGs

The kernel we introduce in this section enlarges the feature space of the ST kernel, with a modest increase in computational burden, and is referred to as $ST+$. In Algorithm 1 we define a procedure to compute the explicit feature space representation $\phi(\cdot)$ of $ST+$. Note that this procedure accesses the graph only by means of $\overset{v}{\Delta}$ and $\overset{v}{\Delta}|_l$, moreover if two trees $\overset{v_i}{\Delta}$ and $\overset{v_j}{\Delta}$ are identical, than also all their subtrees are. Thus, if two nodes generates the same $\pi(v_i) = \pi(v_j)$, then $\overset{v_i}{\Delta} = \overset{v_j}{\Delta}$ and $\overset{ch_m[v_i]}{\Delta}|_l = \overset{ch_m[v_j]}{\Delta}|_l$ for each m and l . Thus, by Theorem 3.2 the procedure is well defined also in the presence of non-comparable nodes, since the resulting tree visits are the same. This proves that the kernel is positive semidefinite. The set of features related to the $ST+$ kernel is a superset of the features of ST and a subset of the features of PT [15]. Line 8 of Algorithm 1 depicts a generic feature introduced by $ST+$. Given a node v and an index j , the feature is defined as the subtree $\overset{v}{\Delta}$ where all subtrees rooted at children of v , except for the j -th child, are replaced by a corresponding limited visit of l levels. Notice that the feature actually depends on $v \in V_D$, the index of a child j and a limit l on the depth of the visits. The function $\pi(f)$ returns the index of the feature f in $\phi(\cdot)$. Figure 2 depicts a partial feature space representation of a DAG according to $ST+$. While for the ST kernel there is one feature for each $v \in V_D$, $ST+$ associates at most $(\rho(v) \cdot h) + 1$ features for any $v \in V_D$. For each node $v \in V_D$, for example the node with label **v** highlighted in Figure 2-a, the algorithm inserts the following features:

Algorithm 1 A procedure for computing the features of the $ST+$ kernel.

```

1: Input: an ordered DAG  $D$ , the maximum depth of the visit  $h$ 
2: for each  $v \in V_D$  do
3:    $f = \triangle_v$ 
4:    $\phi_{\pi(f)} = \phi_{\pi(f)} + \lambda^{\frac{|f|}{2}}$  // add the proper subtree rooted at  $v$  as a feature.
5:   // if the feature is first encountered, it is assumed  $\phi_{\pi(f)} = 0$ 
6:   for  $0 \leq l < \min(h, \text{depth}(f))$  do
7:     for  $1 \leq j \leq \rho(v)$  do

```

$f' = \triangle_{ch_1[v]}^l \quad \dots \quad \triangle_{ch_{j-1}[v]}^l \quad \triangle_{ch_j[v]}^l \quad \triangle_{ch_{j+1}[v]}^l \quad \dots \quad \triangle_{ch_{\rho(v)}[v]}^l$

```

8:      $\phi_{\pi(f')} = \phi_{\pi(f')} + \lambda^{\frac{|f'|}{2}}$  // add the subtree  $f'$  as a feature.
9:   end for
10: end for
11: end for
12: end for
13: Output:  $\phi(\cdot)$ , the set of features of  $D$ 

```

1. the proper subtree rooted at v , which in our example is the one in Figure 2-b;
2. given $ch_j[v]$, the subtree composed by:

- v ;
- the proper subtree rooted at the j -th child of v ;
- the subtrees resulting from a visit limited to $1 \leq l \leq h$ levels starting from the other children of v

is added as feature. As l ranges from 0 to h , the features/subtrees from Figure 2-c to Figure 2-e are added.

Recalling that H is the number of nodes in a DAG ODD_G^v , the complexity of Algorithm 1 is $O(Hh^2\rho^2 \log \rho)$. The complexity of the ODD kernel in (2), instantiated with $ST+$ as base kernel is $O(|V_G| \log |V_G|)$, assuming ρ constant.

4.3. A Novel Feature Weighting Scheme

The features associated with many kernels for graphs, including ODD_{ST_h} and ODD_{ST+} , are not independent from each other. They are, instead, organized in a hierarchical structure [29]. Let us consider the ODD_{ST_h} kernel as an example: given any pair t_i, t such that t_i is a subtree of t , if t occurs as a feature for a graph G , then t_i must occur as features as well. As a consequence, sticking to our example, there is a monotonic increasing relationship between the frequencies of the subtree features t_i and the subtree

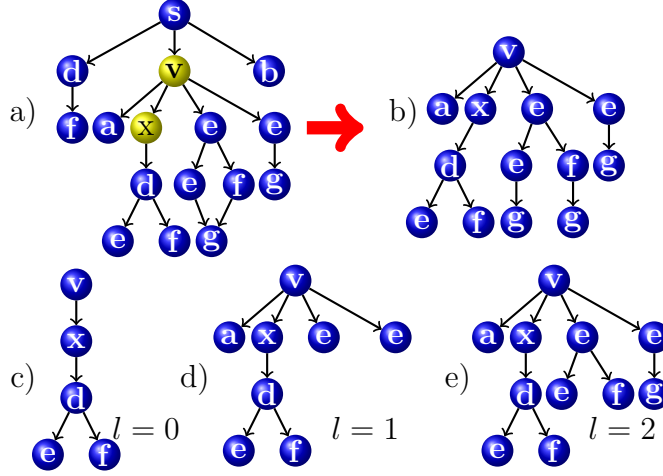


Figure 2: Feature space representation related to the kernel ST_+ : a) an input DAG; b) the proper subtree rooted at the node labelled as **v**; c)-e) given the child **x** of **v**, the features related to visits limited to l levels.

features t . Such relationship is quantified in the upper-left plot of Figure 3, which reports the frequencies of the features generated by the ODD_{ST_h} kernel, for $h \in \{0, \dots, 3\}$, on one of the datasets we will consider in Section 6. The points in the x -axis correspond to features, sorted according to their weights. The y -axis, since $\lambda = 1$, reports the frequencies of the features in the dataset, i.e. the number of times each feature appears in all input graphs. Note that the x -axis is in logarithmic scale. The frequencies are distributed according to a Zipfian distribution, which means that there are very few features with high frequency. Given the structured nature of the feature space, such features are the “simple” ones, i.e. those associated with small sized subtrees, for example single nodes. Any kernel function evaluation will then be highly influenced by such features, which are typically the least discriminative ones. In the case of the ODD_{ST_h} and ODD_{ST_+} kernels, which we recall first decompose the graph into a set of DAGs, the difference between the frequencies of small-sized and large-sized features is even greater since they are extracted from multiple DAGs: the smaller the size of a subtree, the more likely for it to appear in multiple DAGs. The fact that the distribution of weights of the features is particularly skewed, may negatively impact the predictive performance of the kernel since, in principle, we would like to give more emphasis to (i.e. to weight more) bigger, discriminative features with

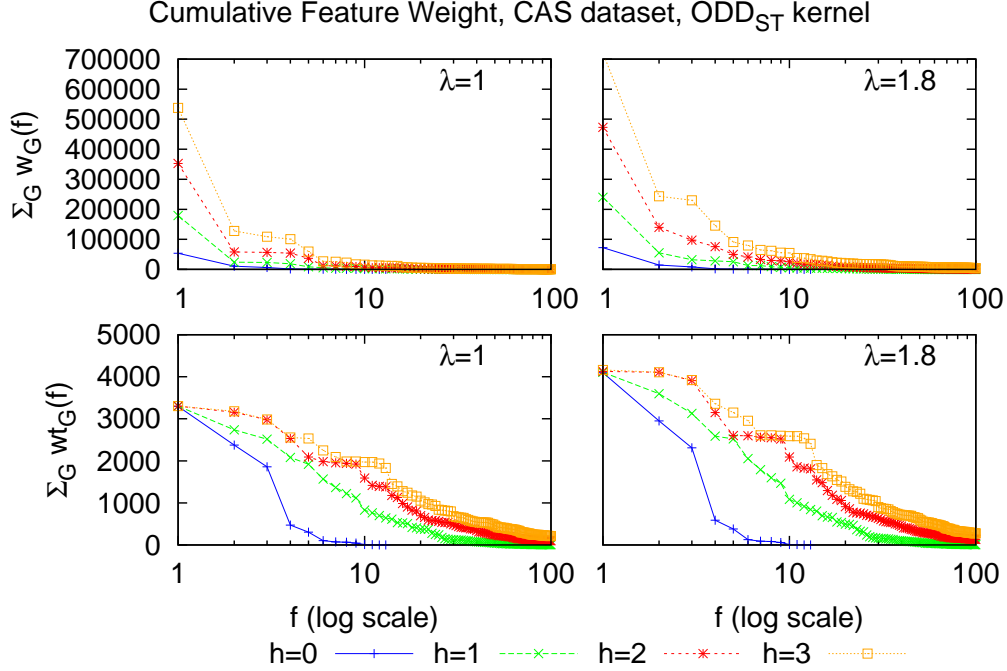


Figure 3: Comparison between the weighting schemes $w_G(f)$ (3) and $wt_G(f)$ (4). On the x axis, in a logarithmic scale, the first 100 features generated by the ODD_{ST_h} kernel for different h values. The y axis reports the cumulative weight of each feature among all the graphs in the dataset.

respect to small ones, that tend to appear in almost all examples, and thus are generally not correlated with the target concept.

One way to tackle this issue is to adopt the weighting scheme explained in Section 4.1, that has been designed specifically for the case of the computation of tree kernels [25]. This scheme has been implemented in the original ODD_{ST_h} kernel formulation, and we maintained it for the proposed ODD_{ST+} kernel: given a graph G , the weight $w_G(f)$ of each feature f (see lines 4 and 9 of Algorithm 1) is computed as

$$w_G(f) = freq_G(f) \cdot \lambda^{\frac{|f|}{2}}, \quad (3)$$

where $freq_G(f)$ is the frequency of the feature f in G . Therefore the contribution to the kernel of the same matching feature (computed via dot product)

in two input graphs G_1 and G_2 is $freq_{G_1}(f) \cdot freq_{G_2}(f) \cdot \lambda^{|f|}$. A value of λ greater than 1 would give more importance to large matching trees. However, the contribution of the less frequent, possibly interesting, small features could be underweighted. The upper-right plot in Figure 3 shows that, with this weighting approach, there are slightly more features with a relatively high weight w.r.t. the case where no weighting scheme is applied (i.e. when $\lambda = 1$). Nonetheless, the distribution is still very skewed.

Another possibility is to define a different weighting scheme, more suited to our approach. As a first step in this direction, we propose to mitigate the contribution of otherwise overweighted features with a different definition¹ of $w_G(f)$, in the following denoted as $wt_G(f)$, i.e.

$$wt_G(f) = \tanh(\lambda^{|f|}) \cdot \tanh(freq_G(f)), \quad (4)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function. Note that the original weighting scheme depends nonlinearly (exponentially) on the size of the feature $|f|$ and linearly on its frequency. The novel scheme we are proposing, on the other hand, depends nonlinearly on both $|f|$ and $freq_G(f)$. In this way, the contribution of each feature is smoothly and non-linearly normalized in the interval $[0, 1]$.

Note that the hyperbolic tangent function is almost linear around zero, and asymptotically tends to one for positive values. This means that the contribution of frequent features is truncated, while the less frequent features are still discriminated since they fall in the linear part of the function. The same is true for the $\lambda^{|f|}$ factor.

The lower plots in Figure 3 reports the weights distribution according to the new wt_G weighting function proposed in (4) with $\lambda = 1$ and $\lambda = 1.8$, respectively. The final result is that the weights are distributed in a smoother way.

The new weighting scheme is applied to the ST kernel, obtaining a variant of the kernel proposed in [27], and to the ST+ kernel first proposed in this paper. Note that this novel weighting scheme is just one possibility among several ones. The key point is that we want to achieve a smoother distribution of the weights associated to the features. The tanh function implements all our desiderata, but any other sigmoidal function can be adopted. Notwith-

¹This is an evolution of the scheme proposed in [26].

standing the heuristic nature of our choice, the experimental results we have obtained on several real world datasets, as reported in Section 6, show that the novel proposed weighting scheme allows to reach statistically significant improvements over state-of-the-art kernels. This seems to confirm that both our intuition on the smoothness of the weight distribution, as well as its implementation via the tanh function, are useful.

5. Related work

Graph data is usually high-dimensional. For this reason, in order to perform learning on graph datasets, there are two possible approaches:

1. applying a preprocessing phase aimed at selecting possibly relevant features;
2. in the context of kernel methods, using tractable kernel functions.

Generally speaking, the methods following the first approach extract frequent patterns, build a vectorial representation of the graphs according to such patterns and then apply a kernel method. When the kernel method is an SVM, the approach is referred to as SVM with frequent pattern mining (freqSVM). The techniques for extracting the features include Gaston [30], Correlated Pattern Mining (CPM) [31], MOLFEA [32]. Saigo et al. [33] proposed gBoost, a boosting method that progressively collects informative (according to the target output) patterns.

The second approach includes a set of kernel functions for graphs. The Marginalized Graph Kernel (MGK) considers common walks as features [34] (the work has been extended in order to make it more efficient and effective in [35]). Informally, this kernel is defined as the expected value of a kernel over all possible pairs of label sequences generated by random walks on two graphs. The worst case time complexity of the algorithm presented in [36] is $O(|V_G|^3)$.

The Shortest Path Kernel associates a feature to each pair of nodes of one graph. The value of the feature is the length of the shortest path between the corresponding nodes in the graph [37]. The complexity of the kernel is $O(|V|^4)$. Being the Shortest Path Kernel based on paths, it can be represented as an instance of (2). We do not report experimental results about this kernel because of its high computational complexity, and its inferior results compared to other state of the art kernels on many of the datasets considered in this paper [24, 38].

In [39] it is described an effective method for computing path based kernels. First a graph is decomposed into a set of trees of totally t nodes. Then the Burrows-Wheeler transform is employed for fast and space-efficient enumeration of paths. The complexity of the kernel is $O(t \log t^\epsilon)$, with $\epsilon < 1$. The *graphlet* kernel [40] counts all types of matching subgraphs of small size k (e.g. $k = 3, 4$ or 5). There are efficient schemes for computing this kernel, but they are applicable only on unlabeled graphs. For the labeled case, the computational complexity of this kernel is $O(n^k)$. In the experimental section of this paper, we considered the Graphlet kernel instantiated with $k = 3$, that will be referred as 3-Graphlet kernel.

The Weisfeiler-Lehman Fast Subtree kernel (FS) counts the number of identical subtree patterns obtained by subtree-walks up to height h [24, 38]. The complexity of the kernel is $O(|E|h)$. While being fast to compute, the kernel may lack of expressiveness for some tasks given that the number of non-zero features generated by one graph is at most $|V|h$. Note that the subtree-walks extracted by the kernel differ from the tree structures extracted by the kernels proposed in Section 4: in FS a node usually appears multiple times in the same subtree-walk, while in the ODD kernel only DAG nodes which have multiple incoming edges appear multiple times in the extracted tree structures. Such difference makes the Weisfeiler-Lehman Fast Subtree kernel not reproducible from (2); a discussion on the differences between the feature spaces induced by the Weisfeiler-Lehman Fast Subtree and the ODD_{ST_h} kernels can be found in [27]. Moreover, the features of the FS kernel are subtree-walks, while specific features (as explained in Sections 4.1 and 4.2) are extracted from the tree-visits obtained from the ODD_{ST_h} and ODD_{ST+} kernels.

Costa and De Grave [21] extended the Fast Subtree Kernel by computing exact matches between pairs of subgraphs with controlled size and distance. Their kernel, named Neighborhood Subgraph Pairwise Distance Kernel (NSPDK), has $O(|V||V_h||E_h|\log|E_h|)$ time complexity, where $|V_h|$ and $|E_h|$ are the number of nodes and the number of edges of the subgraph obtained by a breadth-first visit of depth h . The authors state that, for small values of the subgraph size and distance, the complexity of the kernel becomes in practice linear.

The Weisfeiler-Lehman Shortest path Kernel proposed in [38] is similar in spirit to the NSPDK kernel. Indeed, it considers pairs of subtree patterns and their distance. However it does not limit the maximum distance between the considered patterns, resulting in a computational complexity of $O(n^4)$.

Kernel	Complexity
RW [34]	$O(V ^3)$
SP [37]	$O(V ^4)$
WL-SP [38]	$O(V ^4)$
3-Graphlet [40]	$O(V ^3)$
Treelet [41]	$O(V \rho^5)$
FS [24, 38]	$O(E h)^*$
NSPDK [21]	$O(V)^{*,**}$
ODD _{ST} [27]	$O(V \log V)^*$
ODD _{ST+}	$O(V \log V)^*$

Table 1: Computational complexity of the Shortest Path, the 3-Graphlet, the fast Subtree, the NSPDK, the ODD_{ST} and ODD_{ST+} kernels. *: considering ρ constant; **: with high constants.

Mahé and Vert [23] described a graph kernel based on extracting tree patterns from the graph. The difference with the approach of this paper is that the tree patterns are obtained as result of walks on the graph, i.e. the same node can appear more than once in the same tree pattern. The complexity of the kernel is $O(|V_1||V_2|h\rho^{2\rho})$, where h is the depth of the visit. Finally, [41] proposed the treelet kernel, based on frequent pattern mining of tree-substructures. The kernel implementation considers subtrees with a maximum of 6 nodes, and its computational complexity is $O(n\rho^5)$. Table 1 summarizes the computational complexity of some of the kernels cited in this section, and the ones proposed in this paper. Moreover, just to give an idea about how many features are generated by a graph kernel on a real-world dataset, in Figure 4 we have reported the number of different features generated on a chemical dataset (NCI1) by the most efficient aforementioned kernels.

6. Experimental results

6.1. Experiments on common benchmark graph datasets

The experimental assessment of the proposed kernels has been performed on a total of eight datasets. The first six datasets involve chemo and bioinformatics data: CAS², CPDB [32], AIDS [4], NCI1, NCI109 [3] and GDD [2].

²<http://www.cheminformatics.org/datasets/bursi>

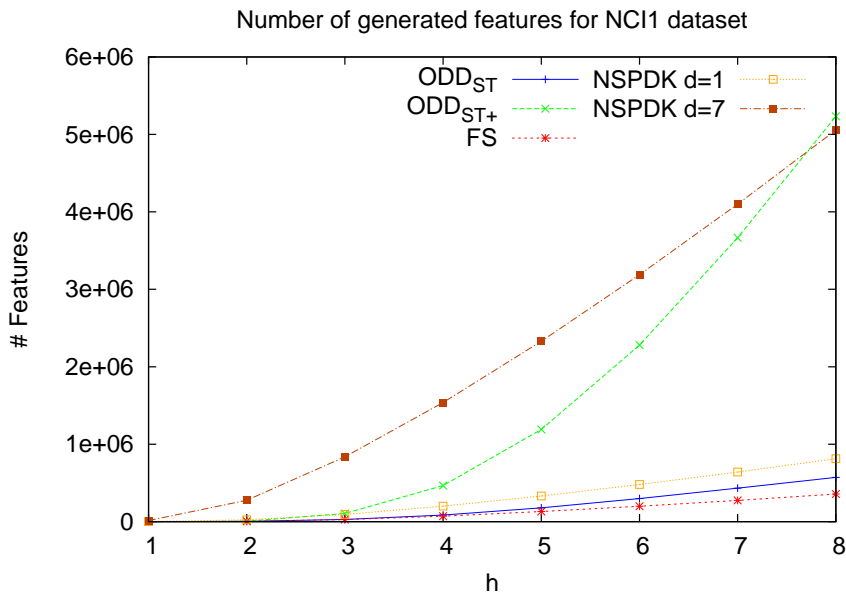


Figure 4: Number of features generated by the ODD_{ST_h} , ODD_{ST_+} , FS and NSPDK kernels on the NCI1 dataset as a function of their parameter h .

The first five datasets involve chemical compounds and represent binary classification problems. The nodes are labeled according to the atom type and the edges represent the bonds. GDD is a dataset composed by proteins represented as graphs, where the nodes of the graphs represent amino acids and two nodes are connected by an edge if they are less than 6Å apart. Moreover, we adopted from [42] two real-world image datasets: MSRC9-class and MSRC21-class³. Each image is represented by its conditional Markov random field graph enriched with semantic labels, and the task is scene classification. Both the datasets are multi-class single-label classification problems. For our experiments, we adopted a SVM classifier [43]. For the multi-class problems, we adopted a one-vs-one scheme. We compare the predictive abilities of the ODD_{ST_+} kernel and the two proposed variants $ODD_{ST_h}^{\text{TANH}}$ and $ODD_{ST_+}^{\text{TANH}}$ to the original ODD_{ST_h} kernel [27], the Fast Subtree Kernel (FS) [24] and the Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [21]. Moreover, we also report the performances of the p -random walk kernel, that is a kernel

³<http://research.microsoft.com/en-us/projects/ObjectClassRecognition/>

Dataset	graphs	pos(%)	avg nodes	avg edges
CAS	4337	55.36	29.9	30.9
CPDB	684	49.85	14.1	14.6
AIDS	1503	28.07	58.9	61.4
NCI1	4110	50.04	29.9	32.3
NCI109	4127	50.37	29.7	32.1
GDD	1178	58.65	284.3	2862.6
MSRC_9	221	multi-class	40.6	97.9
MSRC_21	563	multi-class	77.5	198.3
NCI123	40952	4.76	26.8	28.9
NCL_AIDS	42682	3.52	45.7	47.7

Table 2: Statistics of CAS, CPDB, AIDS, NCI1 , NCI109, GDD, MSRC_9, MSRC_21, NCI123 and NCL_AIDS datasets: number of graphs, percentage of positive examples, average number of atoms, average number of edges.

that compares random walks up to length p in two graphs (special case of [34] and [35]) as representative for the family of kernels based on random walks, and the *graphlet* kernel [40]. Note that the complexity of the *graphlet* kernel (when applied to labeled graphs) is exponential in the size k of the *graphlet*. Because of that, following [38], we restricted our experimentation to a value of k that allows for an efficient computation of the kernel, i.e. $k = 3$.

The experiments are performed using a *nested* 10-fold cross validation: for each of the 10 folds another *inner* 10-fold cross validation, in which we select the best parameters for that particular fold, is performed. All the experiments have been repeated 10 times using different splits for the cross validation, and the average results (with standard deviation) are reported. For all the experiments, the values of the parameters of the ODD_{ST_h} and ODD_{ST+} kernels, including their variants using *tanh*, have been restricted to: $\lambda = \{0.1, 0.2, \dots, 2.0\}$, $h = \{1, 2, \dots, 10\}$. For the Fast Subtree kernel the only parameter $h = \{1, 2, \dots, 10\}$ is optimized. For the NSPDK, the parameters $h = \{1, 2, \dots, 8\}$ and $d = \{1, 2, \dots, 7\}$ are optimized. Finally, for the p -random walk kernel we selected $p = \{1, 2, \dots, 10\}$, and for the *graphlet* kernel we considered only the graphlets of size 3, as mentioned above. A 10x10 CV test with confidence level 95% (and 10 degrees of freedom) has been executed between each pair of kernels on all datasets [44]. In the following the term significant will refer to this statistical test. Table 3 reports the

<i>Kernel</i>	CAS	CPDB	AIDS	NCI1
<i>p</i> -random walk	70.16* (8) ±0.20	64.14* (8) ±1.35	73.55* (8) ±0.49	- ±-
Graphlet	71.10* (7) ±0.48	67.36* (7) ±0.96	73.98* (7) ±0.65	69.68* (7) ±0.52
FS	83.32* (6) ±0.37	76.36 (5) ±1.48	82.02 (5) ±0.4	84.41 (4) ±0.49
NSPDK	83.60* (2) ±0.34	76.99 (1) ±1.15	82.71 (1) ±0.66	83.45 (5) ±0.43
ODD _{ST_h}	83.34* (4) ±0.31	76.44 (4) ±0.62	81.51 (6) ±0.74	82.10* (6) ±0.42
ODD _{ST_h} ^{TANH}	83.40* (3) ±0.41	76.56 (3) ±0.97	82.51 (3) ±0.52	84.57 (3) ±0.43
ODD _{ST+}	83.90 (1) ±0.33	76.30 (6) ±0.23	82.06 (4) ±0.70	84.97 (1) ±0.47
ODD _{ST+} ^{TANH}	83.33* (5) ±0.34	76.74 (2) ±1.81	82.54 (2) ±0.75	84.81 (2) ±0.41
<i>Kernel</i>	GDD	NCI109	MSRC_9	MSRC_21
<i>p</i> -random walk	- ±-	- ±-	67.01* (7) ±2.22	18.88* (8) ±1.4
3-Graphlet	74.92 (6) ±1.40	68.07* (7) ±0.31	60.83* (8) ±2.0	19.66* (7) ±0.96
FS	75.46 (3) ±0.98	85.02 (1) ±0.44	89.26* (6) ±0.82	89.87 (6) ±0.71
NSPDK	74.09 (7) ±0.91	84.17 (2) ±0.33	89.48* (4) ±1.0	90.24 (3) ±0.49
ODD _{ST_h}	75.27 (5) ±0.68	81.91* (6) ±0.42	90.80 (3) ±1.10	89.92 (5) ±0.73
ODD _{ST_h} ^{TANH}	76.09 (1) ±0.85	83.68 (4) ±0.39	94.39 (1) ±1.21	92.60 (1) ±0.45
ODD _{ST+}	75.33 (4) ±0.81	83.08* (5) ±0.49	89.33* (5) ±1.2	89.94 (4) ±0.80
ODD _{ST+} ^{TANH}	75.52 (2) ±0.88	83.93 (3) ±0.42	92.99 (2) ±1.26	91.74 (2) ±0.77

Table 3: Average accuracy results \pm standard deviation in nested 10-fold cross validation for the *p*-random walk, the Graphlet, the Fast Subtree, the Neighborhood Subgraph Pairwise Distance, the ODD_{ST_h} , the $ODD_{ST_h}^{TANH}$, the ODD_{ST+} and the ODD_{ST+}^{TANH} kernels on CAS, CPDB, AIDS, NCI1, GDD, NCI109, MSRC_9 and MSRC_21 datasets. The rank of the kernel is reported between brackets. The symbol * denotes the kernels whose performance difference with respect to the top-ranked kernel is statistically significant.

average accuracies and the rankings obtained by the different kernels on the considered datasets. The symbol * in Table 3 denotes, for each dataset, the kernels whose performance difference with respect to the top-ranked kernel is statistically significant.

Let us now focus on the experimental results obtained for the six chemical datasets. The kernels $ODD_{ST_h}^{TANH}$, ODD_{ST+} , ODD_{ST+}^{TANH} together have best accuracy on three out of six datasets, and the second best accuracy on two others. On the datasets in which the FS and NSPDK kernels perform better than the ODD ones, i.e. CPDB, AIDS and NCI109, the performance difference, at least with respect to the best performing ODD kernel, is never significant. Note that ODD_{ST+} performs significantly better than NSPDK and FS on the CAS dataset. The variant employing the hyperbolic tangent is always useful for the ST kernel, making it the best performing kernel on GDD, and is able to boost the accuracy performance of ODD_{ST+} on AIDS, CPDB, GDD and NCI109 datasets. The generally good results of the ODD kernels, with respect to FS and NSPDK, may be attributed to the fact that they have associated a large feature space, which makes them more adaptable to different tasks. Note that the execution of p -random walk kernel did not complete in 4 days for NCI1, NCI109 and GDD datasets, so the results are missing.

Let us now focus on the image datasets (MSRC_9 and MSRC_21). On these datasets, the baselines FS, NSPDK, ODD_{ST_h} kernels and the proposed ODD_{ST+} kernel show very similar performances. On these datasets, the introduction of the hyperbolic tangent weighting scheme is very beneficial. Both $ODD_{ST_h}^{TANH}$ and ODD_{ST+}^{TANH} performs better than all the baselines, with the former being the best performing kernel on both datasets.

The p -random walk kernel and the *graphlet* kernel show poor performances on these datasets. We argue that this is because they are the only ones among the considered kernels that do not consider all the neighbors of a node as a feature.

Figures 5 and 6 report the computational times required by the ODD_{ST_h} , ODD_{ST+} , NSPDK and the FS kernels as a function of the parameter h determining the size of the considered substructures on the NCI1 and CAS datasets, respectively.

All the experiments are performed on a PC with two Quad-Core AMD Opteron(tm) 2378 Processors and 64GB of RAM. The proposed kernels have been implemented in C++. In addition, we implemented a fast version of

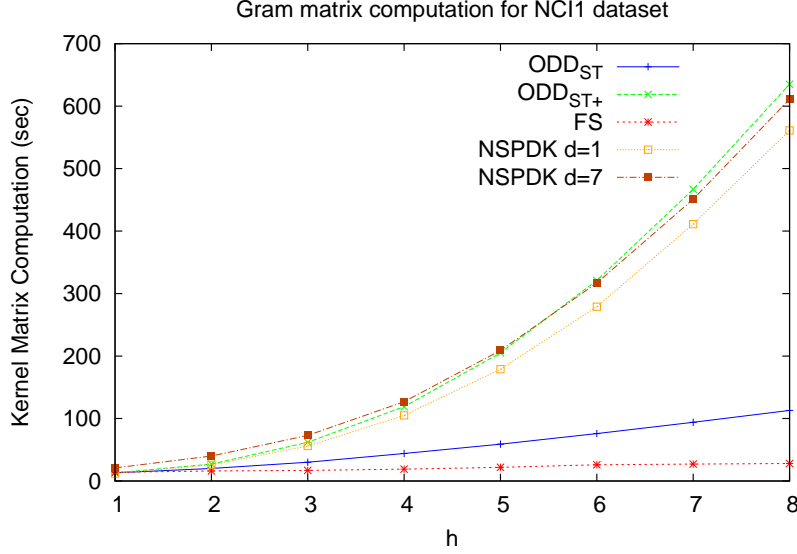


Figure 5: Time needed to compute the kernel matrix for the ODD-ST_h , ODD-ST_h^+ , the NSPDK and the FS kernels, as a function of their parameter h , on NCI1.

the FS kernel in C++. All these kernels adopt an hashing function, similar in spirit to [45]. As for the p -random walk and *graphlet* kernels, we adopted a publicly available Matlab implementation⁴. Thus, the times for the p -random walk and the *graphlet* kernels are reported just for a qualitative comparison. The time needed to compute the kernel matrix for the ODD_{ST^+} kernel increases roughly linearly with respect to the parameter h for both datasets. As expected the constant factors are higher than the ones of the ODD_{ST_h} , but the ODD_{ST^+} is faster than (or comparable to) NSPDK. Note that we do not report the computational times for $\text{ODD}_{ST_h}^{\text{TANH}}$ and $\text{ODD}_{ST^+}^{\text{TANH}}$ since their computational requirements are basically the same as the corresponding base kernels: the computation of the novel weight function does not add a significant computational burden.

Moreover, in Table 4 we report the average computational time for a single fold with the optimal parameters on the four largest datasets: CAS,

⁴<http://www.di.ens.fr/~shervashidze/code.html>

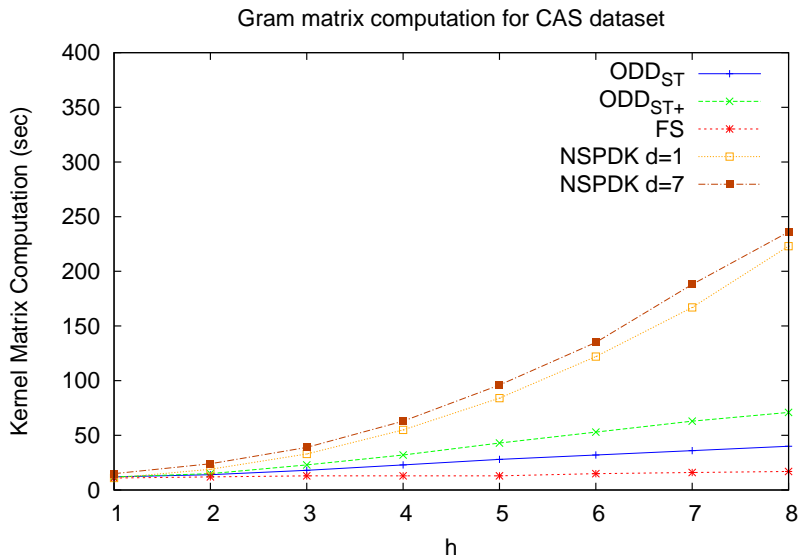


Figure 6: Time needed to compute the kernel matrix for the ODD_{ST_h} , ODD_{ST+_h} , the NSPDK and the FS kernels, as a function of their parameter h , on CAS dataset.

AIDS, NCI1, GDD. The parameters influencing the speed of the kernel are reported between brackets. In this case, we reported the times corresponding to all the considered kernels. Even when comparing the executions related to the optimal parameters, ODD_{ST+} is faster or comparable to NSPDK and ODD_{ST_h} is faster or comparable to FS.

6.2. Experiments on full NCI datasets

In this set of experiments, we analyze how the proposed kernels and the competitors scale up with bigger datasets. We considered two datasets, NCI123 and NCI_AIDS, each one with more than 40,000 examples (see Table 2).

In NCI123⁵ the growth inhibition of the MOLT-4 human Leukemia tumor cell line is measured as a screen for anti-cancer activity. For each compound an activity score of $-\text{LogGI50}$ is measured, where GI50 is the concentration of the compound required for 50% inhibition of tumor growth. A compound

⁵<http://pubchem.ncbi.nlm.nih.gov/bioassay/123>

Kernel	CAS	AIDS	NCI1	GDD
Graphlet	58''	54''	133''	1715''
p -random walk	76 h	35 h	—	—
	(h=7)	(h=8)	(h=—)	(h=—)
FS	13''	5''	28''	17''
	(h=3)	(h=9)	(h=8)	(h=1)
NSPDK	24''	217''	192''	395''
	(h=2, d=6)	(h=8,d=6)	(h=5,d=4)	(h=2,d=6)
ODD_{ST_h}	18''	56''	44''	29''
	(h=3)	(h=7)	(h=4)	(h=1)
$ODD_{ST_h}^{\text{TANH}}$	47''	51''	110''	246''
	(h=5)	(h=6)	(h=6)	(h=2)
ODD_{ST+}	32''	111''	205''	199''
	(h=4)	(h=8)	(h=1)	(h=1)
ODD_{ST+}^{TANH}	179''	61''	165''	541''
	(h=8)	(h=5)	(h=4)	(h=2)

Table 4: Average time required for computing the kernel matrix for the p -random walk, the Graphlet, the Fast Subtree, the Neighborhood Subgraph Pairwise Distance, the ODD_{ST_h} , the $ODD_{ST_h}^{\text{TANH}}$, the ODD_{ST+} and the ODD_{ST+}^{TANH} kernels on CAS, AIDS, NCI1 and GDD datasets with the optimal kernel parameters (reported between brackets).

is classified as active (positive class) or inactive (negative class) if the activity score is, respectively, above or below a specified threshold. The dataset is composed by 40,952 examples. NCI_AIDS⁶ is an anti-HIV database that contains 42,682 molecules, experimentally detected to protect (confirmed active), moderately protect (confirmed moderate) or not protect (inactive) the CEM cells from HIV-1 infection. From these classes we derived a binary classification problem, i.e. distinguishing inactive from confirmed and moderately protective molecules.

Since these two datasets are unbalanced, for this set of experiments we adopted the Area Under the Receiver Operating Characteristic curve (AUROC or AUC) as performance measure, since it is suited for unbalanced datasets. The experimental setup in this case is different w.r.t. the one pre-

⁶<http://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data>

<i>Kernel</i>	NCI123	NCI_AIDS
Graphlet	54.93* (7) ± 0.24	67.74* (7) ± 0.15
FS	61.08* (6) ± 0.34	83.73* (5) ± 0.17
NSPDK	62.45 (3) ± 0.39	83.80* (3) ± 0.23
ODD_{ST_h}	62.11 (4) ± 0.30	83.77* (4) ± 0.22
$ODD_{ST_h}^{TANH}$	62.76 (2) ± 0.21	85.56 (2) ± 0.23
ODD_{ST+}	61.70* (5) ± 0.36	83.36* (6) ± 0.30
ODD_{ST+}^{TANH}	63.20 (1) ± 0.29	85.64 (1) ± 0.15

Table 5: Average AUC results \pm standard deviation in nested 10-fold cross validation for the Graphlet, the Fast Subtree, the Neighborhood Sub-graph Pairwise Distance, the ODD_{ST_h} , the ODD_{ST+} , the $ODD_{ST_h}^{TANH}$ and the ODD_{ST+}^{TANH} kernels obtained on NCI123 and NCI_AIDS datasets. The rank of the kernel is reported between brackets. The symbol * denotes the kernels whose performance difference with respect to the top-ranked kernel is statistically significant.

sented in Section 6.1. Indeed, when the number of examples is large, computing the Gram matrix is unfeasible. In this case, for each considered kernel configuration, we computed the explicit features (memorized in a sparse format) associated to each example. With this explicit feature representation, it is possible to train a linear SVM⁷. Note that the computed solution is equivalent to the one that can be found by a C-SVM applied to the kernel matrix generated by the graph kernel. However, in this way it is possible to handle very large datasets in a reasonable amount of time. A 10x10 CV test with confidence level 95% (and 10 degrees of freedom) has been executed between each pair of kernels on the two datasets [44]. Table 5 reports the AUC results obtained, for the two considered datasets, by kernels for which it is possible to generate the explicit feature space representation of input examples. The

⁷In our implementation we adopted *Liblinear* [46].

combination of the techniques proposed in the paper, ST_+ and \tanh , leads to best performances on both datasets. The performance difference between ODD_{ST_+} e $ODD_{ST_+}^{TANH}$ is statistically significant on both datasets. The use of \tanh yields statistically significant improved performances for $ODD_{ST_+}^{TANH}$ on NCI_AIDS with respect to all other kernels except $ODD_{ST_h}^{TANH}$.

Figure 7 reports the average computational time required to perform the learning procedure for a fixed kernel, as a function of the h parameter, for the NCI123 and NCI_AIDS datasets. This procedure comprehends the feature generation step, and the training phase of the linear SVM model. We decided to report the overall times here because the run-times of linear SVM depends on the characteristics of the kernel, and thus comparing only the feature generation part would not be fair. With the considered learning procedure, the number of non-zero features generated by the kernel influences the total run-time. Indeed, the FS kernel is the fastest one, being the one that generates the smallest number of features. The time required by the training procedure grows almost linearly for ODD_{ST_h} , $ODD_{ST_h}^{TANH}$ and ODD_{ST_+} , while it grows more than linearly for $ODD_{ST_+}^{TANH}$. Note, however, that $ODD_{ST_+}^{TANH}$ is still faster than NSPDK. It is interesting to note that NSPDK with $d = 1$ is slower than NSPDK with $d = 7$ on NCI123, even if the latter has a larger feature space. In this case, probably the former kernel is less discriminative and thus the corresponding optimization problem that the linear SVM must solve is more difficult.

Table 6 reports the computational time required to compute the different kernels with the optimal parameters obtained by a 10-fold cross validation. Note that higher computational times generally corresponds to higher values for the optimal h parameter.

On the considered datasets, higher AUC corresponds to higher computational times for the respective kernel. It is interesting to analyze the relationship between AUC values and running times for non-optimal parameters, i.e. to understand which kernel is the most convenient if there is a strict time constraint to comply to. Figure 8 plots the performances of the different kernels with respect to the time required to perform the training procedure, for NCI123 and NCI_AIDS datasets. In NCI123 dataset, $ODD_{ST_h}^{TANH}$ and $ODD_{ST_+}^{TANH}$ have the highest points in the plot starting from approximatively a runtime of 400 seconds. Below that computational time, the NSPDK is the best performing kernel. On the other hand, on NCI_AIDS dataset, $ODD_{ST_h}^{TANH}$ and $ODD_{ST_+}^{TANH}$ are the better performing kernels for almost every time threshold.

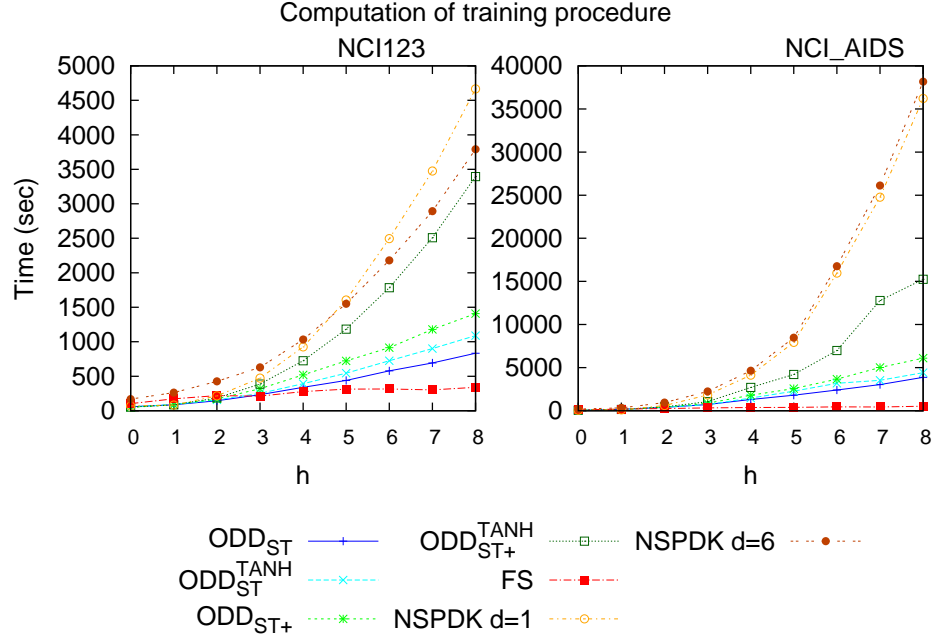


Figure 7: Time needed to perform all the training procedure, as a function of h , for all the considered kernels on NCI123 (left) and NCI_AIDS (right) datasets.

7. Conclusions and future works

The contribution of this paper is twofold. First, we propose a novel instance of the ODD graph kernel based on a novel tree kernel, $ST+$. This constitutes an example of how the generality of the framework can potentially lead to the definition of novel graph kernels that can improve the state-of-the-art. Second, we define a novel, non-linear, feature weighting scheme for the ODD kernels, that can in principle be applied to any graph kernel with an explicit feature space representation. As a future work, we plan to apply this and other weighting schemes also to other state-of-the-art graph kernels. The experimental results show that the proposed kernels have state of the art performances on six benchmark graph datasets from bioinformatics, and on two graph datasets for image classification. Moreover, experiments on two large graph datasets show that our approach is able to scale up to real-world sized datasets.

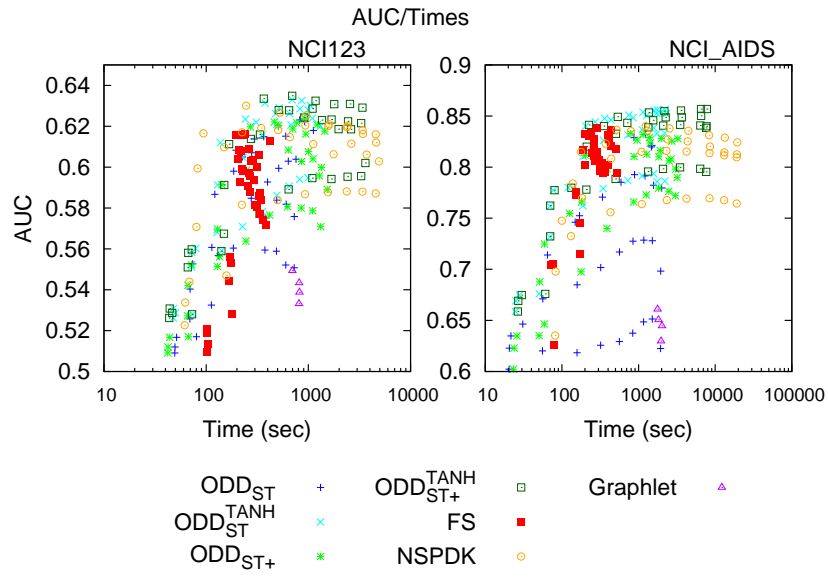


Figure 8: Relationship between the AUC (obtained in 10-fold cross validation) and the time needed to perform all the training procedure. A point is reported for each h and C parameters combination, for all the considered kernels on NCI123 (left) and NCI_AIDS (right) dataset. Note that the x axis is in log scale.

Acknowledgments

This work was supported by the University of Padova under the strategic project *BIOINFOGEN*.

References

- [1] L. Denoyer, P. Gallinari, Report on the XML mining track at INEX 2005 and INEX 2006: categorization and clustering of XML documents, SIGIR Forum 41 (1) (2007) 79–90, ISSN 0163-5840, doi: <http://doi.acm.org/10.1145/1273221.1273230>.
- [2] P. D. Dobson, A. J. Doig, Distinguishing Enzyme Structures from Non-enzymes Without Alignments, Journal of Molecular Biology 330 (4) (2003) 771–783, ISSN 0022-2836, doi:10.1016/S0022-2836(03)00628-4.
- [3] N. Wale, I. Watson, G. Karypis, Comparison of descriptor spaces for chemical compound retrieval and classification, Knowledge and Information Systems 14 (3) (2008) 347–375, ISSN 0219-1377.
- [4] O. S. Weislow, R. Kiser, D. L. Fine, J. Bader, R. H. Shoemaker, M. R. Boyd, New soluble-formazan assay for HIV-1 cytopathic effects: application to high-flux screening of synthetic and natural products for AIDS-antiviral activity., Journal of the National Cancer Institute 81 (8) (1989) 577–586, ISSN 0027-8874.
- [5] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, New York, NY, USA, ISBN 0521813972, 2004.
- [6] G. Simões, H. Galhardas, D. Matos, A Labeled Graph Kernel for Relationship Extraction, in: CoRR, URL <http://arxiv.org/abs/1302.4874>, 2013.
- [7] G. D. Vries, Graph Kernels for Task 1 and 2 of the Linked Data Data-Mining Challenge 2013, in: DMoLD, 2013.
- [8] L. Wang, H. Sahbi, Directed Acyclic Graph Kernels for Action Recognition, 2013 IEEE International Conference on Computer Vision (2013) 3168–3175doi:10.1109/ICCV.2013.393.

- [9] S. Bleik, M. Mishra, J. Huan, M. Song, Text categorization of biomedical data sets using graph kernels and a controlled vocabulary., *IEEE/ACM transactions on computational biology and bioinformatics* / IEEE, ACM 10 (5) (2013) 1211–7, ISSN 1557-9964, doi:10.1109/TCBB.2013.16.
- [10] K. Kundu, F. Costa, R. Backofen, A graph kernel approach for alignment-free domain-peptide interaction prediction with an application to human SH3 domains., *Bioinformatics* (Oxford, England) 29 (13) (2013) i335–43, ISSN 1367-4811, doi:10.1093/bioinformatics/btt220.
- [11] N. Cesa-Bianchi, A. Conconi, C. Gentile, A Second-Order Perceptron Algorithm, *SIAM Journal on Computing* 34 (3) (2005) 640–668.
- [12] D. Haussler, Convolution Kernels on Discrete Structures, Tech. Rep., Department of Computer Science, University of California at Santa Cruz, 1999.
- [13] M. Collins, N. Duffy, New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, 263–270, 2002.
- [14] S. V. N. Vishwanathan, A. J. Smola, Fast kernels for string and tree matching, in: *Advances in Neural Information Processing Systems 15*, MIT Press, 569–576, 2003.
- [15] A. Moschitti, Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees, in: *ECML*, vol. 4212 of *Lecture Notes in Computer Science*, ISBN 3-540-45375-X, 318–329, 2006.
- [16] F. Aioli, G. Da San Martino, A. Sperduti, Route kernels for trees, in: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, ACM Press, New York, New York, USA, ISBN 9781605585161, 17–24, doi:10.1145/1553374.1553377, 2009.
- [17] F. Aioli, G. Da San Martino, A. Sperduti, Extending Tree Kernels with Topological Information, *ICANN* 6791 (2011) 142–149.
- [18] D. Bacciu, A. Micheli, A. Sperduti, A Generative Multiset Kernel for Structured Data, in: A. E. P. Villa, W. Duch, P. Érdi, F. Masulli,

- G. Palm (Eds.), ICANN (1), vol. 7552 of *Lecture Notes in Computer Science*, Springer, ISBN 978-3-642-33268-5, 57–64, 2012.
- [19] T. Gartner, P. Flach, S. Wrobel, T. Gärtner, On Graph Kernels: Hardness Results and Efficient Alternatives, in: B. Schölkopf, M. K. Warmuth (Eds.), Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, vol. 2777 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-40720-1, 129–143, doi:10.1007/b12006, 2003.
 - [20] L. Schietgat, F. Costa, J. Ramon, L. De Raedt, Maximum common subgraph mining: a fast and effective approach towards feature generation, in: 7th International Workshop on Mining and Learning with Graphs, 1–3, 2009.
 - [21] F. Costa, K. De Grave, Fast neighborhood subgraph pairwise distance kernel, in: J. F. Joachims, Thorsten (Eds.), Proceedings of the 27th International Conference on Machine Learning (ICML-10), Omnipress, 255–262, 2010.
 - [22] F. Suard, a. Rakotomamonjy, a. Bensrhair, Kernel on bag of paths for measuring similarity of shapes, European Symposium on Artificial Neural Networks (2007) 1–6.
 - [23] P. Mahé, J. Vert, Graph kernels based on tree patterns for molecules, *Machine Learning* 75 (1) (2009) 3–35.
 - [24] N. Shervashidze, K. M. Borgwardt, Fast subtree kernels on graphs, in: NIPS, 1660–1668, 2009.
 - [25] M. Collins, N. Duffy, Convolution Kernels for Natural Language, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), NIPS, MIT Press, 625–632, 2001.
 - [26] G. Da San Martino, N. Navarin, A. Sperduti, Exploiting the ODD framework to define a novel effective graph kernel., in: proceedings of the 23th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

- [27] G. Da San Martino, N. Navarin, A. Sperduti, A Tree-Based Kernel for Graphs, in: Proceedings of the Twelfth SIAM International Conference on Data Mining, 975–986, 2012.
- [28] G. Da San Martino, N. Navarin, A. Sperduti, A memory efficient graph kernel, in: the 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012.
- [29] P. Yanardag, S. V. N. Vishwanathan, The Structurally Smoothed Graphlet Kernel, arXiv .
- [30] J. Kazius, S. Nijssen, J. Kok, T. Back, A. P. Ijzerman, Substructure Mining Using Elaborate Chemical Representation, J. Chem. Inf. Model. 46 (2) (2006) 597–605.
- [31] B. Bringmann, A. Zimmermann, L. D. Raedt, S. Nijssen, Don’t Be Afraid of Simpler Patterns, in: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), PKDD, vol. 4213 of *Lecture Notes in Computer Science*, Springer, ISBN 3-540-45374-1, 55–66, 2006.
- [32] C. Helma, T. Cramer, S. Kramer, L. De Raedt, Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds, Journal of Chemical Information and Computer Sciences 44 (4) (2004) 1402–1411, ISSN 0095-2338, doi:10.1021/ci034254q.
- [33] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, K. Tsuda, gBoost: a mathematical programming approach to graph classification and regression., Machine Learning (2009) 69–89.
- [34] H. Kashima, K. Tsuda, A. Inokuchi, Marginalized Kernels Between Labeled Graphs., in: T. Fawcett, N. Mishra (Eds.), ICML, AAAI Press, ISBN 1-57735-189-4, 321–328, 2003.
- [35] P. Mahé, N. Ueda, T. Akutsu, J. Perret, J. Vert, Extensions of marginalized graph kernels, in: Proceedings of the twenty-first international conference on Machine learning, ACM, Banff, Alberta, Canada, 70, 2004.
- [36] S. V. N. Vishwanathan, K. M. Borgwardt, N. N. Schraudolph, Fast Computation of Graph Kernels, in: NIPS, 1449–1456, 2006.

- [37] K. M. Borgwardt, H.-P. Kriegel, Shortest-Path Kernels on Graphs, in: Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society, ISBN 0-7695-2278-5, 74–81, 2005.
- [38] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, K. M. Borgwardt, Weisfeiler-Lehman Graph Kernels, *Journal of Machine Learning Research* 12 (2011) 2539–2561.
- [39] M. Heinonen, J. Rousu, N. Välimäki, V. Mäkinen, Efficient Path Kernels for Reaction Function Prediction, in: BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, 202–207, 2012.
- [40] N. Shervashidze, K. Mehlhorn, T. H. Petri, S. V. N. Vishwanathan, K. M. Borgwardt, T. H. Petri, K. Mehlhorn, K. M. Borgwardt, Efficient graphlet kernels for large graph comparison, in: D. van Dyk, M. Welling (Eds.), Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 5 of *JMLR: Workshop and Conference Proceedings*, PASCAL EPrints (United Kingdom), CSAIL, Clearwater Beach, Florida, USA, ISBN 1938-7228, 488–495, 2009.
- [41] B. Gaüzère, P.-A. Grenier, L. Brun, D. Villemin, Treelet kernel incorporating cyclic, stereo and inter pattern information in chemoinformatics, *Pattern Recognition* 48 (2) (2015) 356–367, ISSN 0031-3203.
- [42] M. Neumann, N. Patricia, R. Garnett, K. Kersting, Efficient Graph Kernels by Randomization, in: P. A. Flach, T. De Bie, N. Cristianini (Eds.), ECML PKDD, vol. 7523 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-33459-7, 378–393, 2012.
- [43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830, ISSN 15324435.
- [44] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, NY, USA, ISBN 0521196000, 9780521196000, 2011.

- [45] K. Kersting, M. Mladenov, R. Garnett, M. Grohe, Power Iterated Color Refinement, in: 28th AAAI Conference on Artificial Intelligence, 1904–1910, 2013.
- [46] LIBLINEAR: A library for large linear classification, The Journal of Machine Learning 9 (2008) (2008) 1871–1874.

<i>Kernel</i>	NCI123	NCLAIDS
Graphlet	698 (C=0.01)	1772 (C=0.001)
FS	261 (h=4,C=0.1)	237 (h=3,C=0.1)
NSPDK	246 (h=2,d=5,C=1)	1240 (h=3,d=6,C=1)
ODD _{ST_h}	850 (h=7,C=100)	1608 (h=5,C=100)
ODD _{ST_h} ^{TANH}	692 (h=6,C=1)	2219 (h=8,C=1)
ODD _{ST₊}	924 (h=5,C=10)	790 (h=3,C=10)
ODD _{ST₊} ^{TANH}	694 (h=4,C=1)	7739 (h=8,C=1)

Table 6: Time needed to perform all the training procedure with the optimal parameter configuration (reported between brackets) for all the considered kernels on NCI123 and NCLAIDS datasets.